Twitter Data Wrangling Report.

By Temitope Adebiyi.

The data wrangling report entails, gathering of data, assessing data and cleaning the data.

Gathering Data

There were three different pieces of data I gathered from different sources. The first one which is the WeRateDogs Twitter archive, which I downloaded manually from the Udacity project site and uploaded it to the jupyter notebook, and used pandas library to import the data frame.

The second data is the tweet image predictions that is what breed of Dog is present in each tweet according to a neural network. The file was hosted on Udacity's servers and was downloaded programmatically using the Request library.

The third data which consist of tweet;s retweet count and favorite count at minimum and some additional data I found interesting. I had to apply as a developer on twitter to get my Twitter API keys, secrets, and tokens for the project. I used the tweet IDs in the WeRateDogs Twitter archive, queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's set of JSON data in a file called tweet_json.txt. Then I read the txt file line by line into a pandas DataFrame.

Assessing Data

After gathering each of the above pieces of data. I assessed them visually and programmatically for quality and tidiness issue. The quality issues are as below:

- Rows with retweets which is not needed for the project will be dropped.
- Drop irrelevant columns and columns that have incomplete values
- There is issues with the rating denominator there are value not equals to 10
- In the image prediction column of p1, p2, p3 there are ' _' in the values instead of space.
- In the image prediction data frame there should be one column for the best prediction and another column for the best confidence level.
- Rating numerator column has enormous values which should be dropped
- The timestamp column should be converted to datetime data type.
- The dog name column has some incorrect values like none, a, an, the, etc which are not actual names.
- There are dogs that have multiple dog stages

The tidiness issues are as below:

- The dog name should be in one column and the various stages of dog should be in a single column.
- The three data frame should be merge together to form a single data frame.
- Delete the old prediction columns after the getting the most correct prediction

Cleaning Data

I programmatically cleaned the Data Frames, according to the quality issues and tidiness issues listed above and stored the cleaned Data Frames in a file called 'Cleaned data'.